CHAIN OF THOUGHT PROMPTING UNDER STREAMING BATCH: A CASE STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, Large Language Models (LLMs) have demonstrated remarkable capabilities. Chain-of-Thought (CoT) has been proposed as a way of assisting LLMs in performing complex reasoning. However, developing effective prompts can be a challenging and labor-intensive task. Many studies come out of some way to automatically construct CoT from test data. Most of them assume that all test data is visible before testing and only select a small subset to generate rationales, which is an unrealistic assumption. In this paper, we present a case study on how to construct and optimize chain-of-thought prompting using batch data in streaming settings.

1 INTRODUCTION

Large Language Models (LLMs) have shown emergent abilities. Wei et al. (2022a) Wei et al. (2022b) discover Chain-of-Thought (CoT) prompting as a simple and broadly applicable method for enhancing reasoning in language models. Many work Zhou et al. (2022a); Wang et al. (2022b); Shi et al. (2022); Zhang et al. (2023a;b); Wang et al. (2022a); Zhou et al. (2022b) has tried to make further improvements based on CoT.

2 PROBLEM STATEMENT

The problem of *chain-of-thought under streaming* is first proposed by Zhang et al. (2023a). It assumes that the test dataset, denoted as *D*, will be evenly partitioned into m batches, each containing N samples. These batches are fed into a large language model M in a continuous stream-like manner. At each time step T^k for $k = 1, 2, \ldots, m$, M processes one batch of questions $\left\{q_1^{(k)}, q_2^{(k)}, \ldots, q_N^{(k)}\right\}$ in batch B^k using the same prompt P. After processing a batch, rationales generated by M are denoted as $\left\{c_1^{(k)}, c_2^{(k)}, \ldots, c_N^{(k)}\right\}$. A *prompting optimization function* $f(P|(q_1^{(k)}||c_1^{(k)}), (q_2^{(k)}||c_2^{(k)}), \ldots, (q_N^{(k)}||c_N^{(k)}))$[1] is applied to update prompt P before processing the next batch. This allows the model to maintain a coherent chain-of-thought within each batch in the stream, known as the *intra-batch chain-of-thought*.

Function f adopted in Zhang et al. (2023a) is a simple concatenation function by appending all the newly generated question-rationale pairs $(q_1^{(k)}||c_1^{(k)}), (q_2^{(k)}||c_2^{(k)}), \ldots, (q_N^{(k)}||c_N^{(k)})$ to the previous prompting P. However, this approach can quickly reach the maximum input sequence length of the Language Model (e.g. 2048 tokens) and is not particularly scalable or efficient, potentially leading to high levels of redundancy in the prompting and high costs for querying LLM. Therefore, an alternative approach may be needed to address these limitations.

3 PROMPTING OPTIMIZATION FUNCTION

Prompt optimization is a black-box optimization problem, as it can only be evaluated based on the quality of the generated rationales or the correctness of the answers provided by LLM. In this regard, we present our empirical findings on how to design f to update prompting constrained by the

[1] The concatenation of question and rationale is denoted as a question-rationale pair $(q||c)$

limit input sequence length based on two attributes of CoT: **correctness** and **depth**. **Correctness** is a crucial criterion for prompting engineers to update the prompting. Here, we want to ask the question: *Can a valid rationale be replaced with an invalid one without performance drop?* **Depth** refers to the number of reasoning steps which can be reflected by the length of CoT (different steps are separated by comma or \n). A deeper CoT is longer and typically contains more complex rationales, while a shallower CoT is more straightforward and contains fewer reasoning steps. Here, the question is: *Given the similar question, can a deep CoT always be replaced by a shallow CoT?*

4 EXPERIMENTS

Our evaluation is conducted on model `text-davinci-002` from OpenAI. We choose multiple datasets across different tasks including arithmetic reasoning (GSM8K Cobbe et al. (2021), MultiArith Roy & Roth (2015)), commonsense reasoning (StrategyQA Geva et al. (2021)), and symbolic reasoning (Letter Wei et al. (2022b)). We compare our methods with Zero-Shot-CoT Kojima et al. (2022), and Bootstraping Auto-CoT Zhang et al. (2023a). We manually partition each dataset into 10 batches and generate rationales for each batch under the streaming setting. We report the test accuracy for ten batches. The number of samples in each batch is in Appendix A.

4.1 RIGHT OR WRONG COT

To address this question, we substitute ground-truth rationales that produce the correct answer with Zero-Shot-CoT that produce the wrong answer. During the substitution, we ensure that more than 50% of the rationales in the prompt are incorrect. This substitution method is referred to as *Wrong-CoT*. Conversely, we have *Correct-CoT*, which includes only the correct rationales. Figure 1(a) shows the results.

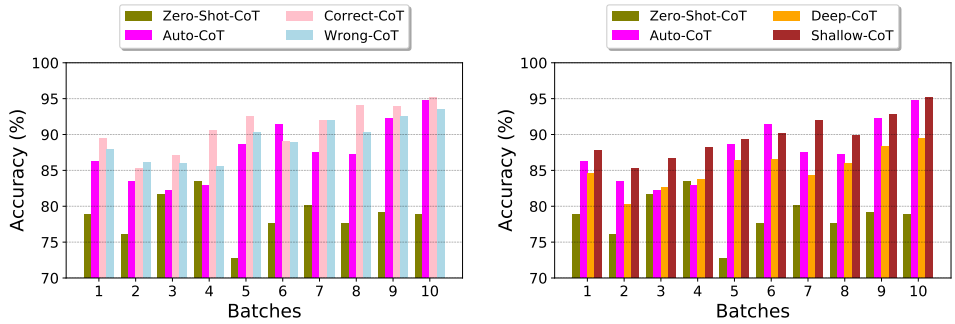

Figure 1: **Left:** accuracy for MultiArith dataset under Correct-CoT and Wrong-CoT. **Right:** accuracy for MultiArith dataset under Deep-CoT and Shallow-CoT with $\xi = 3$.

4.2 DEEP OR SHALLOW COT

In order to address the aforementioned question, we exclusively utilize CoTs that are deemed correct in the prompt. To distinguish between deep CoT and shallow CoT, we have established a simple heuristic parameter denoted as ξ, which is based on the number of \n in the CoT. If the number of \n surpasses ξ, we classify it as a deep CoT. Conversely, if a CoT has fewer \n than ξ, we classify it as a shallow CoT. Figure 1(b) shows the results.

5 CONCLUSION

This paper presents a straightforward case study on how to update and replace prompts in a streaming batch setting. Our findings indicate that incorrect chain-of-thought promptings can be valuable and do not significantly diminish performance. Additionally, promptings that consist of shorter rationales exhibit superior performance compared to those with lengthier rationales.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies, 2021. URL https://arxiv.org/abs/2101.02235.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Subhro Roy and Dan Roth. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1743–1752, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1202. URL https://aclanthology.org/D15-1202.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022b. URL https://arxiv.org/abs/2201.11903.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023a.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023b.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022a.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022b.

A APPENDIX

Table 1: Batch size of four different datasets

	MultiArith	GSM8K	StrategyQA	Letter
Batch Size	60	64	32	81

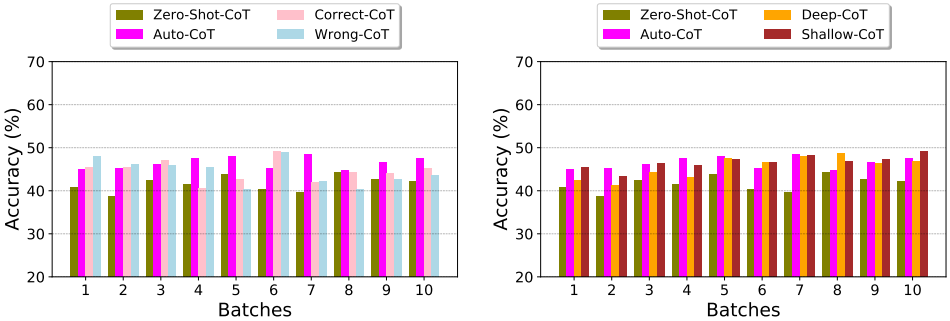

Figure 2: **Left:** accuracy for GSM8K dataset under Correct-CoT and Wrong-CoT. **Right:** accuracy for GSM8K dataset under Deep-CoT and Shallow-CoT with $\xi = 3$.

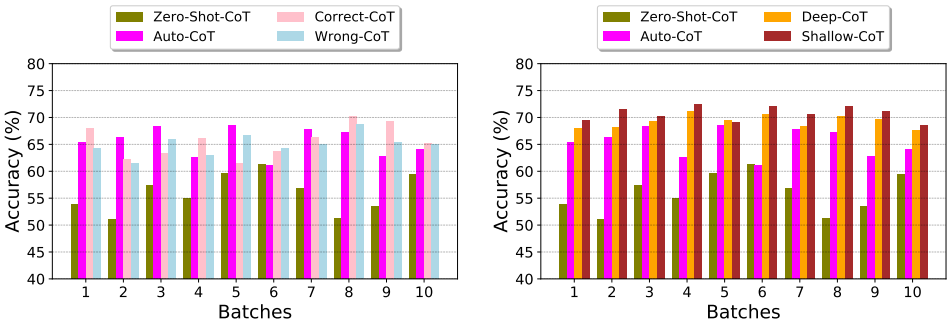

Figure 3: **Left:** accuracy for StrategyQA dataset under Correct-CoT and Wrong-CoT. **Right:** accuracy for StrategyQA dataset under Deep-CoT and Shallow-CoT with $\xi = 3$.

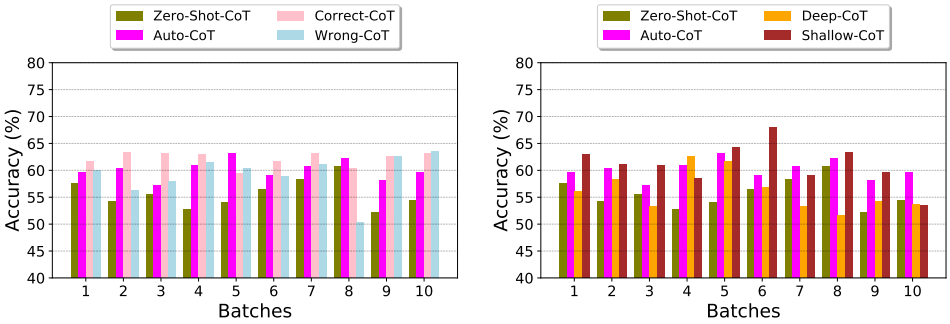

Figure 4: **Left:** accuracy for Letter dataset under Correct-CoT and Wrong-CoT. **Right:** accuracy for Letter dataset under Deep-CoT and Shallow-CoT with $\xi = 4$.